Variational Latent Dependency Learning

Jiawei He* **& Yu Gong***
Simon Fraser University
jha203, gongyug@sfu.ca

Joseph Marino
California Institute of Techology
jmarino@caltech.edu

Greg Mori
Simon Fraser University
mori@cs.sfu.ca

Andreas Lehrmann
andreas.lehrmann@gmail.com

1 Introduction

In this work, we propose a method for learning dependency structures in latent variable models. We discuss a variational end-to-end approach for learning arbitrary directed graph structures introducing minimal complexity overhead. In particular, we introduce a set of binary global variables to gate the latent dependencies. The whole model (including its structure) is jointly optimized with a single stochastic variational inference objective. In our experimental validation, we show that the learned dependency structures contribute to a more accurate representation of the true generative distribution, outperforming several other variants of variational autoencoders.

Variational autoencoders (VAEs) [Kingma and Welling, 2014, Rezende et al., 2014] amortize inference optimization in latent variable models across data examples $\mathbf{x}$ and latent variables $\mathbf{z}$ by parameterizing $q_\phi(\mathbf{z}|\mathbf{x})$ as a separate inference model, then jointly optimizing the model parameters θ and ϕ using evidence lower bound: $\mathcal{L}(\mathbf{x};\theta,\phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))$. However, VAEs are typically implemented with basic graphical structures and simple, unimodal distributions (e.g. Gaussians). Similarly, approximate posteriors often make the mean field assumption, $q_\phi(\mathbf{z}|\mathbf{x}) = \prod_m q_\phi(z_m|\mathbf{x})$. Independence assumptions such as these may be overly restrictive, thereby limiting modeling capabilities.

Incorporating dependency structure among the latent variables is one way to improve expressiveness [Sønderby et al., 2016, He et al., 2018, Marino et al., 2018]. These dependencies provide *empirical* priors, learned priors that are conditioned on other latent variables. With M latent dimensions, the full prior takes the following auto-regressive form: $p_\theta(\mathbf{z}) = \prod_{m=1}^{M} p_\theta(z_m|\mathbf{z}_{\mathrm{pa}(m)})$, where $\mathbf{z}_{\mathrm{pa}(m)}$ denotes the vector of latent variables constituting the parents of z_m. The marginal empirical prior of such models, $p_\theta(z_m) = \int p_\theta(z_m|\mathbf{z}_{\mathrm{pa}(m)})p_\theta(\mathbf{z}_{\mathrm{pa}(m)})\mathrm{d}\mathbf{z}_{\mathrm{pa}(m)}$, can be arbitrarily complex.

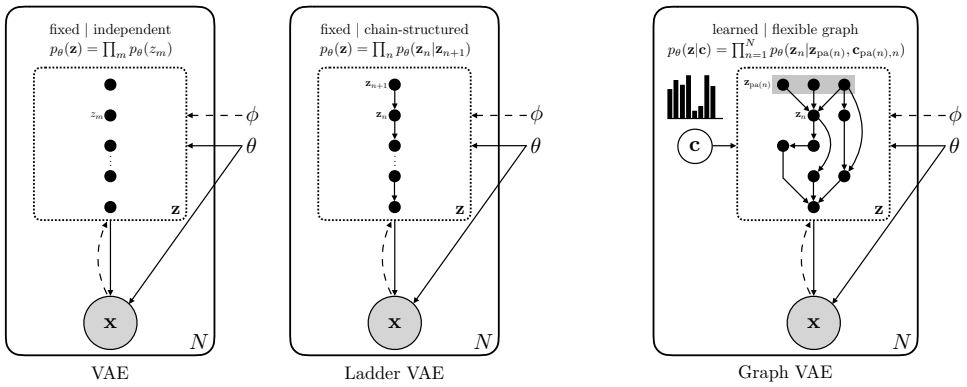
Rather than relying on pre-defined fully-connected structures or chain structures [Sønderby et al., 2016] in previous works, we seek to automatically learn the latent dependency structure as part of the variational optimization process. A comparison of these approaches is visualized in Fig. 1.

2 Variational Optimization of Latent Structures

Given a fixed number of latent dimensions, a finite number of possible dependency structures exists and can be modelled with a fully-connected directed acyclic graph (DAG). Naively, one might assume that a fully-connected DAG is the most general structure, and, therefore, current models should implicitly learn to ignore unnecessary dependencies between latent variables. However, latent variable models are highly prone to local optima, and as we show empirically in experiments, modifying entire dependencies can yield improved performance.

*Equal Contribution.

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.



(a) Traditional Latent Variable Models (b) Proposed Model

Figure 1: **Overview: Model Comparison.** We show the graphical representations of **(a)** traditional latent variable models (VAE, ladder VAE) and **(b)** the proposed graph VAE. Solid lines denote generation, dashed lines denote inference, and the dotted area indicates the latent space governed by variational parameters ϕ and generative parameters θ. Both VAE and ladder VAE use a fixed graph structure with limited expressiveness (VAE: independent; ladder VAE: chain-structured). In contrast, graph VAE jointly optimizes a distribution over latent structures $\mathbf{c}$ and model parameters (ϕ, θ), allowing test-time sampling of a flexible, data-driven latent structure.

To control the dependency structure of the model, we introduce a set of binary global variables, $\mathbf{c} = \{c_{i,j}\}_{i,j}$, which *gate* the latent dependencies. The element $c_{i,j}$ denotes the gate variable from $\mathbf{z}_i$ to $\mathbf{z}_j$ ($i > j$), specifying the presence or absence of this latent dependency. We treat each $c_{i,j}$ as an independent random variable, sampled from a Bernoulli distribution with mean $\mu_{i,j}$, i.e. $c_{i,j} \sim \mathcal{B}(\mu_{i,j})$. We denote the set of these Bernoulli means as μ. With this addition, the model is now expressed as $p_\theta(\mathbf{x}, \mathbf{z}, \mathbf{c}) = p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{c})p_\theta(\mathbf{z}|\mathbf{c})p(\mathbf{c})$, and the prior can now be expressed as $p_\theta(\mathbf{z}|\mathbf{c}) = \prod_{n=1}^N p_\theta(\mathbf{z}_n|\mathbf{z}_{\text{pa}(n)}, \mathbf{c}_{\text{pa}(n),n})$, where $\mathbf{c}_{\text{pa}(n),n}$ denotes the gate variables associated with the dependencies between node $\mathbf{z}_n$ and its parents, $\mathbf{z}_{\text{pa}(n)}$. Note that $\mathbf{z}_{\text{pa}(n)}$ denotes the set of all *possible* parents of node $\mathbf{z}_n$ in the fully-connected DAG, i.e. $\mathbf{z}_{\text{pa}(n)} = \{\mathbf{z}_{n+1}, \ldots, \mathbf{z}_N\}$.

Introducing the dependency gating variables modifies the variational objective, as we must now marginalize over these additional variables. The corresponding lower bound can thus be expressed as

$$\widetilde{\mathcal{L}} = \mathbb{E}_{p(\mathbf{c})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{c})\right] - \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{c}))\right] = \mathbb{E}_{p(\mathbf{c})}\left[\mathcal{L}_\mathbf{c}\right], \tag{1}$$

where $\mathcal{L}_\mathbf{c}$ is the ELBO for a given value of dependency gating variables, $\mathbf{c}$. Thus, $\widetilde{\mathcal{L}}$ can be interpreted as the expected ELBO under the distribution of dependency structures induced by $p(\mathbf{c})$. We form a Monte Carlo estimate of $\widetilde{\mathcal{L}}$ by sampling $\mathbf{c} \sim p(\mathbf{c})$ and evaluating $\mathcal{L}_\mathbf{c}$.

For a given latent dependency structure, gradients for the parameters θ and ϕ can be estimated using Monte Carlo samples and the reparameterization trick. To obtain gradients for the gate means, μ, we make use of recent advances [Jang et al., 2017] in differentiating through discrete operations, allowing us to differentiate through the sampling of the dependency gating variables, $\mathbf{c}$. Specifically, we recast the gating variables using the Gumbel-Softmax estimator, re-expressing $c_{i,j}$ as:

$$c_{i,j} = \frac{\exp((\log(\mu_{i,j}) + \epsilon_1)/\tau)}{\exp((\log(\mu_{i,j}) + \epsilon_1)/\tau) + \exp((\log(1 - \mu_{i,j}) + \epsilon_2)/\tau)}, \tag{2}$$

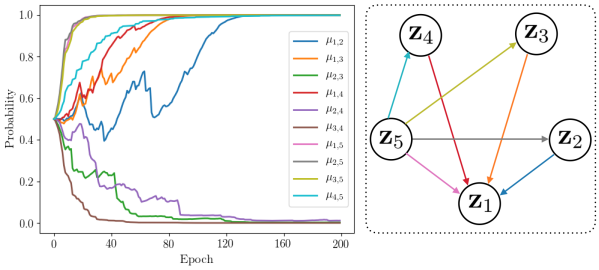
where ϵ_1 and ϵ_2 are i.i.d samples drawn from a Gumbel(0, 1) distribution and τ is a temperature parameter. The Gumbel-Softmax distribution is differentiable for $\tau > 0$, allowing us to estimate the derivative $\frac{\partial c_{i,j}}{\partial \mu_{i,j}}$.

3 Experiments

We evaluate the proposed model and its learned latent dependency structure on three challenging datasets: MNIST [Lecun et al., 1998], Omniglot [Lake et al., 2013], and CIFAR-10 [Krizhevsky,

Dataset	Method								
	MNIST			Omniglot			CIFAR-10		
	LL	KL	ELBO	LL	KL	ELBO	LL	KL	ELBO
VAE	-89.1	29.0	-92.1	-110.9	30.5	-120.4	-6.63	0.110	-6.69
Ladder VAE	-84.8	**24.3**	-87.8	-106.4	**27.9**	-112.5	-6.47	0.082	-6.50
FC VAE	-83.0	28.9	-84.8	-104.8	29.9	-106.6	-6.44	0.077	-6.46
Graph VAE	$\mathbf{-82.1}$	27.8	$\mathbf{-84.1}$	$\mathbf{-103.4}$	29.1	$\mathbf{-105.2}$	$\mathbf{-6.40}$	**0.074**	$\mathbf{-6.41}$

Table 1: **Quantitative Analysis.** We show test-time log-likelihoods (LL), $D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))$ (KL), and ELBO of the proposed graph VAE model (last row) and compare it to baselines on 3 popular datasets.



(a) (b)

Figure 2: **Structure Learning.** In (**a**), we show training of the Bernoulli parameters $\mu_{i,j}$ governing the distribution over graph structures in architecture space. All edges are color-coded and can be located in (**b**), where we show a random sample from the resulting steady-state distribution with the same color scheme.

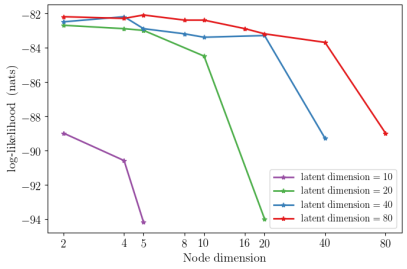


Figure 3: **Ablation Study on MNIST.** For a fixed latent dimension M (color-coded), we report the log-likelihood of all possible factorizations of node dimension N' (x-axis) and number of nodes $N = M/N'$.

2009]. Our baselines consist of classic variational autoencoders as well as its popular variants, including ladder VAEs and VAEs with fully-connected latent dependency structure (FC-VAEs).

An example of the structure learning process on MNIST is visualized in Fig. 2. In Fig. 2a, we show the evolution of the parameters μ of the gating variables $\mathbf{c}$. The network actively drops $3/10$ edges during the learning process, while the remaining parameters eventually converge to 1. A sample of the learned distribution is shown in Fig. 2b. Fig. 3 reports an ablation study on the influence of the total latent dimension M and, for fixed latent dimension, the trade-off between number of nodes N and node dimension $N' = M/N$.

We evaluate the performance of all models using their test-time log-likelihood $\log p_\theta(\mathbf{x})$ (Table 1). All values were estimated using $5,000$ importance-weighted samples. Following standard practice, we report $\log p_\theta(\mathbf{x})$ in *nats* on MNIST/Omniglot and in *bits/input dimension* on CIFAR-10. Our proposed model with a jointly optimized, learned dependency structure consistently outperforms both models with less expressive (VAE, ladder VAE) and more expressive (FC-VAE) predefined structures. To provide further insights into the training objective, Table 1 also reports $D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}))$ and the full ELBO objective (Eq. (1)) of the trained models.

4 Conclusion

We presented a novel method for optimizing variational autoencoders jointly with their latent dependency structures. Our experiments showed that the learned latent dependency structure improves the generative performance of latent variable models. By introducing an additional set of binary structure indicator variables and optimizing a structure-dependent evidence lower bound, our model is able to learn a better representation of the underlying generative distribution than baselines with a predetermined structure.

References

Jiawei He, Andreas Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. Probabilistic video generation using holistic attribute control. In *European Conference on Computer Vision*, 2018.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations*, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Brenden M Lake, Ruslan R Salakhutdinov, and Josh Tenenbaum. One-shot learning by inverting a compositional causal process. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2526–2534. Curran Associates, Inc., 2013.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. ISSN 0018-9219. doi: 10.1109/5.726791.

Joseph Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In *International Conference on Machine Learning*, pages 3400–3409, 2018.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in neural information processing systems*, pages 3738–3746, 2016.